1

Establishing Trust

Artificial intelligence is the study of machines that exhibit traits associated with a human mind such as perception, learning, reasoning, planning, and problem solving. Although it had a prior history under different names (e.g. cybernetics and automata studies), we may consider the genesis of the field of artificial intelligence to be the Dartmouth Summer Research Project on Artificial Intelligence in the summer of 1956. Soon thereafter, the field split into two camps: one focused on symbolic systems, problem solving, psychology, performance, and serial architectures, and the other focused on continuous systems, pattern recognition, neuroscience, learning, and parallel architectures.[1] This book is primarily focused on the second of the two partitions of artificial intelligence, namely machine learning.

The term *machine learning* was popularized in Arthur Samuel's description of his computer system that could play checkers,[2] not because it was explicitly programmed to do so, but because it learned from the experiences of previous games. In general, machine learning is the study of algorithms that take data and information from observations and interactions as input and *generalize* from specific inputs to exhibit traits of human thought. Generalization is a process by which specific examples are abstracted to more encompassing concepts or decision rules.

One can subdivide machine learning into three main categories: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. In supervised learning, the input data includes observations and labels; the labels represent some sort of true outcome or common human practice in reacting to the observation. In unsupervised learning, the input data includes only observations. In reinforcement learning, the inputs are interactions with the real world and rewards accrued through those actions rather than a fixed dataset.

[1] Allen Newell. "Intellectual Issues in the History of Artificial Intelligence." In: *The Study of Information: Interdisciplinary Messages*. Ed. by Fritz Machlup and Una Mansfield. New York, New York, USA: John Wiley & Sons, 1983, pp. 187–294.

[2] A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers." In: *IBM Journal of Research and Development* 3.3 (Jul. 1959), pp. 210–229.

The applications of machine learning may be divided into three broad categories: (1) cyber-physical systems, (2) decision sciences, and (3) data products. Cyber-physical systems are engineered systems that integrate computational algorithms and physical components, e.g. surgical robots, self-driving cars, and the smart grid. Decision sciences applications use machine learning to aid people in making important decisions and informing strategy, e.g. pretrial detention, medical treatment, and loan approval. Data products applications are the use of machine learning to automate informational products, e.g. web advertising placement and media recommendation. These settings vary widely in terms of their interaction with people, the scale of data, the time scale of operation and consequence, and the severity of consequences. Trustworthy machine learning is important in all three application categories, but is typically more pronounced in the first two categories: cyber-physical systems and decision sciences. In data products applications, trustworthy machine learning contributes to a functioning non-violent society.

Just a few years ago, the example applications in all of the categories would have been unheard of. In recent years, however, machine learning has achieved superlative performance on several narrowly-defined tasks across domains (often surpassing the abilities of human experts on those same tasks) and invaded the popular imagination due to the confluence of three factors: data, algorithms, and computation. The amount of data that is captured digitally and thus available to machine learning algorithms has increased exponentially. Algorithms such as deep neural networks have been developed to generalize well from that data. Computational paradigms such as graphical processing units and cloud computing have allowed machine learning algorithms to tractably learn from very large datasets.

The end result is that machine learning has become a general purpose technology that can be used in many different application domains for many different uses. Like other general purpose technologies before it,[3] such as the domestication of plants, the wheel, and electricity, machine learning is starting to remake all parts of society. In some parts of the world, machine learning already has an incipient role in every part of our lives, including health and wellness, law and order, commerce, entertainment, finance, human capital management, communication, transportation, and philanthropy.

Despite artificial intelligence's promise to reshape different sectors, there has not yet been wide adoption of the technology except in certain pockets such as electronic commerce and media. Like other general purpose technologies, there are many short-term costs to the changes required in infrastructure, organizations, and human capital.[4] In particular, it is difficult for many businesses to collect and curate data from disparate sources. Importantly, corporations do not trust artificial intelligence and machine learning in critical enterprise workflows because of a lack of transparency into the inner workings and a potential lack of reliability. For example, a recent study of business

[3] List of general purpose technologies: domestication of plants, domestication of animals, smelting of ore, wheel, writing, bronze, iron, waterwheel, three-masted sailing ship, printing, steam engine, factory system, railway, iron steamship, internal combustion engine, electricity, motor vehicle, airplane, mass production, computer, lean production, internet, biotechnology, nanotechnology. Richard G. Lipsey, Kenneth I. Carlaw, and Clifford T. Bekar. *Economic Transformations*. Oxford, England, UK: Oxford University Press, 2005.

[4] Brian Bergstein. "This Is Why AI Has Yet to Reshape Most Businesses." In: *MIT Technology Review* (Feb. 2019). URL: https://www.technologyreview.com/s/612897/this-is-why-ai-has-yet-to-reshape-most-businesses.

decision makers found that only 21% of them have a high level of trust in different types of analytics;[5] the number is likely smaller for machine learning, which is a part of analytics in business parlance.

> "A decision aid, no matter how sophisticated or 'intelligent' it may be, may be rejected by a decision maker who does not trust it, and so its potential benefits to system performance will be lost."
>
> —Bonnie M. Muir, psychologist at University of Toronto

This book is being written at a juncture in time when there is a lot of enthusiasm for machine learning. It is also a time when many societies are reckoning with social justice. Many claim that it is the beginning of the age of artificial intelligence, but others are afraid of impending calamity. The technology is poised to graduate from the experimental sandboxes of academic and industrial laboratories to truly widespread adoption across domains, but only if the gap in trust can be overcome.

I restrain from attempting to capture the zeitgeist of the age, but provide a concise and self-contained treatment of the technical aspects of machine learning. The goal is not to mesmerize you, but to get you to think things through.[6] There is a particular focus on mechanisms for increasing the trustworthiness of machine learning systems. As you'll discover throughout the journey, there is no single best approach toward trustworthy machine learning applicable across all applications and domains. Thus, the text focuses on helping you develop the thought process for weighing the different considerations rather than giving you a clear-cut prescription or recipe to follow. Toward this end, I provide an operational definition of trust in the next section and use it as a guide on our conceptual development of trustworthy machine learning. I tend to present evergreen concepts rather than specific tools and tricks that may soon become dated.

1.1 Defining Trust

What is trust and how do we operationalize it for machine learning?

> "What is trust? I could give you a dictionary definition, but you know it when you feel it. Trust happens when leaders are transparent, candid, and keep their word. It's that simple."
>
> —Jack Welch, CEO of General Electric

It is not enough to simply be satisfied by: 'you know it when you feel it.' The concept of trust is defined and studied in many different fields including philosophy, psychology, sociology, economics, and organizational management. Trust is the relationship between a *trustor* and a *trustee*: the trustor trusts the trustee. A definition of trust from organizational management is particularly appealing and

[5] Maria Korolov. "Explainable AI: Bringing Trust to Business AI Adoption." In: *CIO* (Sep. 2019). URL: https://www.cio.com/article/3440071/explainable-ai-bringing-trust-to-business-ai-adoption.html.

[6] The curious reader should research the etymology of the word 'mesmerize.'

relevant for defining trust in machine learning because machine learning systems in high-stakes applications are typically used within organizational settings. *Trust is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.*[7] This definition can be put into practice as a foundation for desiderata of machine learning systems.

1.1.1 Trusted vs. Trustworthy

Embedded within this definition is the idea that the trustee has certain properties that make it *trustworthy*, i.e. the qualities by which the trustor can expect the trustee to perform the important action referred to in the definition of trust. Being trustworthy does not automatically imply that the trustee is trusted. The trustor must consciously make a decision to be vulnerable to the trustee based on its trustworthiness and other factors including cognitive biases of the trustor. Understandably, potential trustors who are already vulnerable as members of marginalized groups may not want to become even more vulnerable. A system may not be trusted no matter how worthy of trust it is.

> "The toughest thing about the power of trust is that it's very difficult to build and very easy to destroy."
>
> —Thomas J. Watson, Sr., CEO of IBM

Moreover, the trustor's expectation of the trustee can evolve over time, even if the trustworthiness of the trustee remains constant. A typical dynamic of increasing trust over time begins with the trustor's expectation of performance being based on (1) the *predictability* of individual acts, moves onto (2) expectation based on *dependability* captured in summary statistics, finally culminating in (3) the trustor's expectation of performance based on *faith* that dependability will continue in the future.[8] Predictability could arise from some sort of understanding of the trustee by the trustor (for example their motivations or their decision-making procedure) or by low variance in the trustee's behavior. The expectation referred to in dependability is the usual notion of expectation in probability and statistics.

In much of the literature on the topic, both the trustor and the trustee are people. For our purposes, however, an end-user or other person is the trustor and the machine learning system is the trustee. Although the specifics may differ, there are not many differences between a trustworthy person and a trustworthy machine learning system. However, the final trust of the trustor, subject to cognitive biases, may be quite different for a human trustee and machine trustee depending on the task.[9]

1.1.2 Attributes of Trustworthiness

Building upon the above definition of trust and trustworthiness, you can list many different attributes of a trustworthy person: availability, competence, consistency, discreetness, fairness, integrity, loyalty,

[7] Roger C. Mayer, James H. Davis, and F. David Schoorman. "An Integrative Model of Organizational Trust." In: *Academy of Management Review* 20.3 (Jul. 1995), pp. 709–734.

[8] John K. Rempel, John G. Holmes, and Mark P. Zanna. "Trust in Close Relationships." In: *Journal of Personality and Social Psychology* 49.1 (Jul. 1985), pp. 95–112.

[9] Min Kyung Lee. "Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management." In: *Big Data & Society* 5.1 (Jan.–Jun. 2018).

openness, promise fulfilment, and receptivity to name a few.[10] Similarly, you can list several attributes of a trustworthy information system, such as: correctness, privacy, reliability, safety, security, and survivability.[11] The 2019 International Conference on Machine Learning (ICML) listed the following topics under trustworthy machine learning: adversarial examples, causality, fairness, interpretability, privacy-preserving statistics and machine learning, and robust statistics and machine learning. The European Commission's High Level Expert Group on Artificial Intelligence listed the following attributes: lawful, ethical, and robust (both technically and socially).

Such long and disparate lists give us some sense of what people deem to be trustworthy characteristics, but are difficult to use as anything but a rough guide. However, we can distill these attributes into a set of *separable* sub-domains that provide an organizing framework for trustworthiness. Several pieces of work converge onto a nearly identical set of four such separable attributes; a selected listing is provided in Table 1.1. The first three rows of Table 1.1 are attributes of trustworthy people. The last two rows are attributes of trustworthy artificial intelligence. Importantly, through separability, it is implied that each of the qualities is conceptually different and we can examine each of them in isolation of each other.

Table 1.1. *Attributes of trustworthy people and artificial intelligence systems.*

	Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
trustworthy people	Mishra[12]	competent	reliable	open	concerned
	Maister et al.[13]	credibility	reliability	intimacy	low self-orientation
	Sucher and Gupta[14]	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others' interests as well as its own
trustworthy artificial intelligence	Toreini et al.[15]	ability	integrity	predictability	benevolence
	Ashoori and Weisz[16]	technical competence	reliability	understandability	personal attachment

[10] Graham Dietz and Deanne N. Den Hartog. "Measuring Trust Inside Organisations." In: *Personnel Review* 35.5 (Sep. 2006), pp. 557–588.

[11] Fred B. Schneider, ed. *Trust in Cyberspace*. Washington, DC, USA: National Academy Press, 1999.

[12] Aneil K. Mishra. "Organizational Responses to Crisis: The Centrality of Trust." In: *Trust in Organizations*. Ed. by Roderick M. Kramer and Thomas Tyler. Newbury Park, California, USA: Sage, 1996, pp. 261–287.

[13] David H. Maister, Charles H. Green, and Robert M. Galford. *The Trusted Advisor*. New York, New York, USA: Touchstone, 2000.

[14] Sandra J. Sucher and Shalene Gupta. "The Trust Crisis." In: *Harvard Business Review* (Jul. 2019). URL: https://hbr.org/cover-story/2019/07/the-trust- crisis.

[15] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. "The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain, Jan. 2020, pp. 272–283.

[16] Maryam Ashoori and Justin D. Weisz. "In AI We Trust? Factors That Influence Trustworthiness of AI-Infused Decision-Making Processes." arXiv:1912.02675, 2019.

1.1.3 Mapping Trustworthy Attributes to Machine Learning

Interpreting the attributes of trustworthiness from the table in the context of machine learning is the primary thread of this book. In particular, we take Attribute 1 (competence) to be basic performance such as the accuracy of a machine learning model. Good performance, appropriately quantified based on the specifics of the problem and application,[17] is a necessity to be used in any real-world task.

We take Attribute 2 to include the reliability, safety, security and fairness of machine learning models and systems. Machine learning systems need to maintain good and correct performance across varying operating conditions. Different conditions could come from natural changes in the world or from malevolent or benevolent human-induced changes.

We take Attribute 3 to consist of various aspects of openness and human interaction with the machine learning system. This includes communication from the machine to the human through comprehensibility of models by people as well as transparency into overall machine learning system pipelines and lifecycles. It also includes communication from the human to the machine to supply personal and societal desires and values.

We take Attribute 4 (selflessness) to be the alignment of the machine learning system's purpose with a society's wants. The creation and development of machine learning systems is not independent of its creators. It is possible for machine learning development to go in a dystopian direction, but it is also possible for machine learning development to be intertwined with matters of societal concern and applications for social good, especially if the most vulnerable members of society are empowered to use machine learning to meet their own goals.

Although each of the four attributes are conceptually distinct, they may have complex interrelationships. We return to this point later in the book, especially in Chapter 14. There, we describe relationships among the different attributes (some are tradeoffs, some are not) that policymakers must reason about to decide a system's intended operations.

We use the following working definition of trustworthy machine learning in the remainder of the book. **A trustworthy machine learning system is one that has sufficient:**

1. **basic performance,**
2. **reliability,**
3. **human interaction, and**
4. **aligned purpose.**

We keep the focus on making machine learning systems worthy of trust rather than touching on other (possibly duplicitous) ways of making them trusted.

1.2 Organization of the Book

The organization of the book closely follows the four attributes in the definition of trustworthy machine learning. I am purposefully mindful in developing the concepts slowly rather than jumping ahead quickly to the later topics that may be what are needed in immediate practice. This is because

[17] Kiri L. Wagstaff. "Machine Learning that Matters." In: *Proceedings of the International Conference on Machine Learning*. Edinburgh, Scotland, UK, Jun.–Jul. 2012, pp. 521–528.

the process of creating trustworthy machine learning systems, given the high consequence of considerations like safety and reliability, should also be done in a thoughtful manner without overzealous haste. Taking shortcuts can come back and bite you.

> "Slow down and let your System 2 take control."[18]
>
> —Daniel Kahneman, behavioral economist at Princeton University
>
> "Worry about rhythm rather than speed."
>
> —Danil Mikhailov, executive director of data.org

Highlighted in Figure 1.1, the remainder of Part 1 discusses the book's limitations and works through a couple of preliminary topics that are important for understanding the concepts of trustworthy machine learning: the personas and lifecycle of developing machine learning systems in practice, and quantifying the concept of safety in terms of uncertainty.

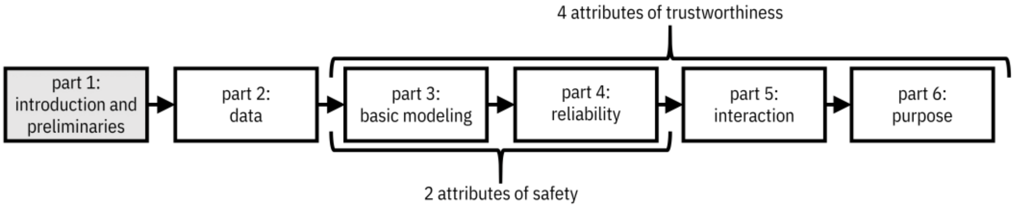

Figure 1.1. *Organization of the book. This first part focuses on introducing the topic of trustworthy machine learning and covers a few preliminary topics.* Accessible caption. A flow diagram from left to right with six boxes: part 1: introduction and preliminaries; part 2: data; part 3: basic modeling; part 4: reliability; part 5: interaction; part 6: purpose. Part 1 is highlighted. Parts 3–4 are labeled as attributes of safety. Parts 3–6 are labeled as attributes of trustworthiness.

Part 2 is a discussion of data, the prerequisite for doing machine learning. In addition to providing a short overview of different data modalities and sources, the part touches on three topics relevant for trustworthy machine learning: biases, consent, and privacy.

Part 3 relates to the first attribute of trustworthy machine learning: basic performance. It describes optimal detection theory and different formulations of supervised machine learning. It teaches several different learning algorithms such as discriminant analysis, naïve Bayes, k-nearest neighbor, decision

[18] Kahneman and Tversky described two ways in which the brain forms thoughts, which they call 'System 1' and 'System 2.' System 1 is fast, automatic, emotional, stereotypic and consciousness. System 2 is slow, effortful, logical, calculating, and conscious. Please engage the 'System 2' parts of your thought processes and be deliberate when you develop trustworthy machine learning systems.

trees and forests, logistic regression, support vector machines, and neural networks. The part concludes with methods for causal discovery and causal inference.

Part 4 is about the second attribute of trustworthy machine learning: reliability. This attribute is discussed through three specific topics: distribution shift, fairness, and adversarial robustness. The descriptions of these topics not only define the problems, but also provide solutions for detecting and mitigating the problems.

Part 5 is about the third attribute: human interaction with machine learning systems in both directions—understanding the system and giving it instruction. The part begins with interpretability and explainability of models. It moves onto methods for testing and documenting aspects of machine learning algorithms that can then be transparently reported, e.g. through factsheets. The final topic of this part is on the machine eliciting the policies and values of people and society to govern its behavior.

Part 6 discusses the fourth attribute: what those values of people and society may be. It begins by covering the ethics principles assembled by different parties as their paradigms for machine learning. Next, it discusses how the inclusion of creators of machine learning systems with diverse lived experiences broadens the values, goals, and applications of machine learning, leading in some cases to the pursuit of social good through the technology. Finally, it shows how the prevailing paradigm of machine learning in information recommendation platforms leads to filter bubbles and disinformation, and suggests alternatives. The final chapter about platforms is framed in terms of trustworthy institutions, which have different attributes than individual trustworthy people or individual trustworthy machine learning systems.

1.3 Limitations

Machine learning is an increasingly vast topic of study that requires several volumes to properly describe. The elements of trust in machine learning are also now becoming quite vast. In order to keep this book manageable for both me (the author) and you (the reader) it is limited in its depth and coverage of topics. Parts of the book are applicable both to simpler data analysis paradigms that do not involve machine learning and to explicitly programmed computer-based decision support systems, but for the sake of clarity and focus, they are not called out separately.

Significantly, despite trustworthy machine learning being a topic at the intersection of technology and society, the focus is heavily skewed toward technical definitions and methods. I recognize that philosophical, legal, political, sociological, psychological, and economic perspectives may be even more important to understanding, analyzing, and affecting machine learning's role in society than the technical perspective. Nevertheless, these topics are outside the scope of the book. Insights from the field of human-computer interaction are also extremely relevant to trustworthy machine learning; I discuss these to a limited extent at various points in the book, particularly Part 5.

Within machine learning, I focus on supervised learning at the expense of unsupervised and reinforcement learning. I do, however, cover graphical representations of probability and causality as well as their inference. Within supervised learning, the primary focus is on classification problems in which the labels are categorical. Regression, ordinal regression, ranking, anomaly detection, recommendation, survival analysis, and other problems without categorical labels are not the focus. The depth in describing various classification algorithms is limited and focused on high-level concepts rather than more detailed accounts or engineering tricks for using the algorithms.

Several different forms and modalities of data are briefly described in Part 2, such as time series, event streams, graphs, and parsed natural language. However, the primary focus of subsequent chapters is on forms of data represented as feature vectors.[19] Structured, tabular data as well as images are naturally represented as feature vectors. Natural language text is also often represented by a feature vector for further analysis.

An important ongoing direction of machine learning research is transfer learning, a paradigm in which previously learned models are repurposed for new uses and contexts after some amount of fine-tuning with data from the new context. A related concept for causal models is statistical transportability. Nonetheless, this topic is beyond the scope of the book except in passing in a couple of places. Similarly, the concepts of multi-view machine learning and causal data fusion, which involve the modeling of disparate sets of features are not broached. In addition, the paradigm of active learning, in which the labeling of data is done sequentially rather than in batch before modeling, is not discussed in the book.

As a final set of technical limitations, the depth of the mathematics is limited. For example, I do not present the concepts of probability at a depth requiring measure theory. Moreover, I stop at the posing of optimization problems and do not go into specific algorithms for conducting the optimization.[20] Discussions of statistical learning theory, such as generalization bounds, are also limited.

1.4 Positionality Statement

It is highly atypical for a computer science or engineering book to consider the influence of the author's personal experiences and background on its contents. Such a discussion is known as a *reflexivity statement* or *positionality statement* in the social sciences. I do so here since power and privilege play a key role in how machine learning is developed and deployed in the real-world. This recognition is increasing because of a current increase in attention to social justice in different societies. Therefore, it is important to be transparent about me so that you can assess potential biases against marginalized individuals and groups in the contents of the book. I'll evaluate myself using the four dimensions of trustworthiness detailed earlier in the chapter (competence, reliability, interaction, and purpose).

> "Science currently is taught as some objective view from nowhere (a term I learned about from reading feminist studies works), from no one's point of view."
>
> —Timnit Gebru, research scientist at Google

I encourage you, the reader, to create your own positionality statement as you embark on your journey to create trustworthy machine learning systems.

[19] A feature is an individual measurable attribute of an observed phenomenon. Vectors are mathematical objects that can be added together and multiplied by numbers.

[20] Mathematical optimization is the selection of a best element from some set of alternatives based on a desired criterion.

1.4.1 Competence and Credibility

I completed a doctorate in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT). My dissertation included a new kind of supervised machine learning method and a decision-theoretic model of human decision making that quantitatively predicts racial bias. I have been a research staff member at IBM Research – Thomas J. Watson Research Center since 2010 conducting research on statistical signal processing, data mining, and machine learning. The results have been published in various reputed workshops, conferences, journals, and magazines including ICML, the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Learning Representations (ICLR), the ACM Conference on Knowledge Discovery and Data Mining (KDD), the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES), the Journal of Machine Learning Research (JMLR), the IEEE Transactions on Signal Processing, the IEEE Transactions on Information Theory, and the Proceedings of the IEEE. I have defined a large part of the strategy for trustworthy machine learning at IBM Research and a large subset of my own work has been on interpretability, safety, fairness, transparency, value alignment, and social good in machine learning and artificial intelligence.

I have developed real-world solutions that have been deployed in high-stakes applications of machine learning and data science during engagements with IBM business units, various clients of IBM, and social change organizations. I have led teams that developed the comprehensive open source toolkits and resources on fairness, explainability and uncertainty quantification named AI Fairness 360, AI Explainability 360 and Uncertainty Quantification 360, and transitioned some of their capabilities into the IBM Watson Studio product. I have spoken at various industry-oriented meetups and conventions such as the O'Reilly AI Conference, Open Data Science Conference, and IBM Think.

I have been an adjunct faculty member at New York University (NYU) and a guest lecturer in courses at Cornell, Georgetown, NYU, Princeton, Rutgers, and Syracuse. I organized the Workshop on Human Interpretability in Machine Learning at ICML annually from 2016 to 2020 as well as several other workshops and symposia related to trustworthy machine learning. I served as a track chair for the practice and experience track of the 2020 ACM Conference on Fairness, Accountability and Transparency and was a member of the Partnership on AI's Safety-Critical AI expert group.

To compose this book, I am channeling all these past experiences along with the interactions with students, lifelong learners, and colleagues that these experiences have afforded. Of course, I have less depth of knowledge about the topics of some of the chapters than others, but have some level of both practical/applied and theoretical knowledge on all of them.

1.4.2 Reliability and Biases

Reliability stems from the ability to work in different contexts and conditions. I have only had one employer, which limits this ability. Nevertheless, by working at IBM Research and volunteering with DataKind (an organization that helps professional data scientists conduct projects with social change organizations), my applied data science work has engaged with a variety of for-profit corporations, social enterprises, and non-profit organizations on problems in human resources and workforce analytics, health systems and policy, clinical health care, humanitarian response, international development, financial inclusion, and philanthropic decision making. Moreover, my research contributions have been disseminated not only in machine learning research venues, but also statistics, operations research, signal processing, information theory, and information systems venues, as well as the industry-oriented venues I mentioned earlier.

More importantly for trustworthy machine learning, I would like to mention my privileges and personal biases. I was born and raised in the 1980s and 1990s in predominantly white upper middle-class suburbs of Syracuse, a medium-sized city in upstate New York located on the traditional lands of the Onöñda'gaga' people, that is one of the most racially-segregated in the United States. Other places I have lived for periods of three months or longer are Ithaca, Elmsford, Ossining, and Chappaqua in New York; Burlington and Cambridge in Massachusetts; Livermore, California; Ludhiana, New Delhi, and Aligarh in northern India; Manila, Philippines; Paris, France; and Nairobi, Kenya. I am a cis male, second-generation American of South Asian descent. To a large extent, I am an adherent of dharmic religious practices and philosophies. One of my great-great-grandfathers was the first Indian to study at MIT in 1905. My father and his parents lived hand-to-mouth at times, albeit with access to the social capital of their forward caste group. My twin brother, father, and both grandfathers are or were professors of electrical engineering. My mother was a public school teacher. I studied in privileged public schools for my primary and secondary education and an Ivy League university for my undergraduate education. My employer, IBM, is a powerful and influential corporation. As such, I have been highly privileged in understanding paths to academic and professional success and having an enabling social network. Throughout my life, however, I have been a member of a minority group with limited political power. I have had some visibility into hardship beyond the superficial level, but none of this experience has been *lived experience*, where I would not have a chance to leave if I wanted to.

1.4.3 Interaction

I wrote the book with some amount of transparency. While I was writing the first couple of chapters in early 2020, anyone could view them through Overleaf (https://v2.overleaf.com/read/bzbzymggkbzd). After I signed a book contract with Manning Publications, chapters were posted to the Manning Early Access Program as I wrote them, with readers having an opportunity to engage via the Manning liveBook Discussion Forum. After the publisher and I parted ways in September 2021, I posted chapters of the in-progress manuscript to http://www.trustworthymachinelearning.com. I received several useful comments from various individuals throughout the drafting process via email (krvarshn@us.ibm.com), Twitter direct message (@krvarshney), telephone (+1-914-945-1628), and personal meetings. When I completed version 0.9 of the book at the end of December 2021, I posted it at the same site. On January 28, 2022, I convened a panel of five people with lived experiences different from mine to provide their perspectives on the content contained in version 0.9 using a modified Diverse Voices method.[21] An electronic version of this edition of the book will continue to be available at no cost at the same website: http://www.trustworthymachinelearning.com.

[21] Lassana Magassa, Meg Young, and Batya Friedman. "Diverse Voices: A How-To Guide for Facilitating Inclusiveness in Tech Policy." Tech Policy Lab, University of Washington, 2017. The panelists who provided impartial input were Mashael Alzaid, Kenya Andrews, Noah Chasek-Macfoy, Scott Fancher, and Timothy Odonga. As a central part of the Diverse Voices method, they were offered honoraria, which some declined. The funds came from an honorarium I received for participating in an AI Documentation Summit convened by The Data Nutrition Project in January 2022.

1.4.4 Motivation and Values

My motivations begin with family values. The great-great-grandfather I mentioned above returned to India with knowledge of industrial-scale glassmaking from MIT and made social impact by establishing a factory in service of *swaraj*, self-governance in India, and the training of local workers. One of my grandfathers applied his knowledge of systems and control theory to problems in agriculture and also worked toward social justice in India through non-technological means. My other grandfather joined UNESCO to establish engineering colleges in developing Iraq and Thailand. My mother taught science in an inner-city school district's special program for students caught with weapons in their regular middle and high schools.

In the same way, consistent with family values as well as external ethics (*yama*),[22] internal ethics (*niyama*),[23] and the ethos of the American dream, my personal motivation is to advance today's most societally-impactful technology (machine learning), mitigate its harmfulness, apply it to uplift humanity, and train others to do the same. I co-founded the IBM Science for Social Good fellowship program in 2015–2016 and direct it toward these aims.

The reason I wrote this book is many-fold. First, I feel that although many of the topics that are covered in the book, like fairness, explainability, robustness, and transparency are often talked about together, there is no source that unifies them in a coherent thread. With this book, there is such a resource for technologists, developers, and researchers to learn from. Second, I feel that in industry practice, the unbridled success of deep learning has led to too much emphasis on engineers squeezing out a little more accuracy with little conceptual understanding and little regard to considerations beyond accuracy (the other three attributes of trust). The aim of the book is to fill the conceptual understanding gap for the practitioners who wish to do so, especially those working in high-stakes application domains. (Cai and Guo find that many software engineers fundamentally desire guidance on understanding and applying the conceptual underpinnings of machine learning.[24]) The inclusion of considerations beyond predictive accuracy cannot be an afterthought; it must be part of the plan from the beginning of any new project. Third, I would like to empower others who share my values and ethics to pursue a future in which there is a virtuous cycle of research and development in which technology helps society flourish and society helps technology flourish.

1.5 Summary

- Machine learning systems are influencing critical decisions that have consequences to our daily lives, but society lacks trust in them.
- Trustworthiness is composed of four attributes: competence, reliability, openness, and selflessness.

[22] List of *yamas*: *ahiṃsā* (non-harm), *satya* (benevolence and truthfulness), *asteya* (responsibility and non-stealing), *brahmacarya* (good direction of energy), and *aparigraha* (simplicity and generosity).

[23] List of *niyamas*: *śauca* (clarity and purity), *santoṣa* (contentment), *tapas* (sacrifice for others), *svādhyayā* (self-study), and *īsvara-praṇidhāna* (humility and service to something bigger).

[24] Carrie J. Cai and Philip J. Guo. "Software Developers Learning Machine Learning: Motivations, Hurdles, and Desires." In: *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*. Memphis, Tennessee, USA, Oct. 2019, pp. 25–34.

- The book is organized to match this decomposition of the four components of trust.
- Despite my limitations and the limitations of the contents, the book endeavors to develop a conceptual understanding not only of the principles and theory behind how machine learning systems can achieve these goals to become more trustworthy, but also develop the algorithmic and non-algorithmic methods to pursue them in practice.
- By the end of the book, your thought process should naturally be predisposed to including elements of trustworthiness throughout the lifecycle of machine learning solutions you develop.